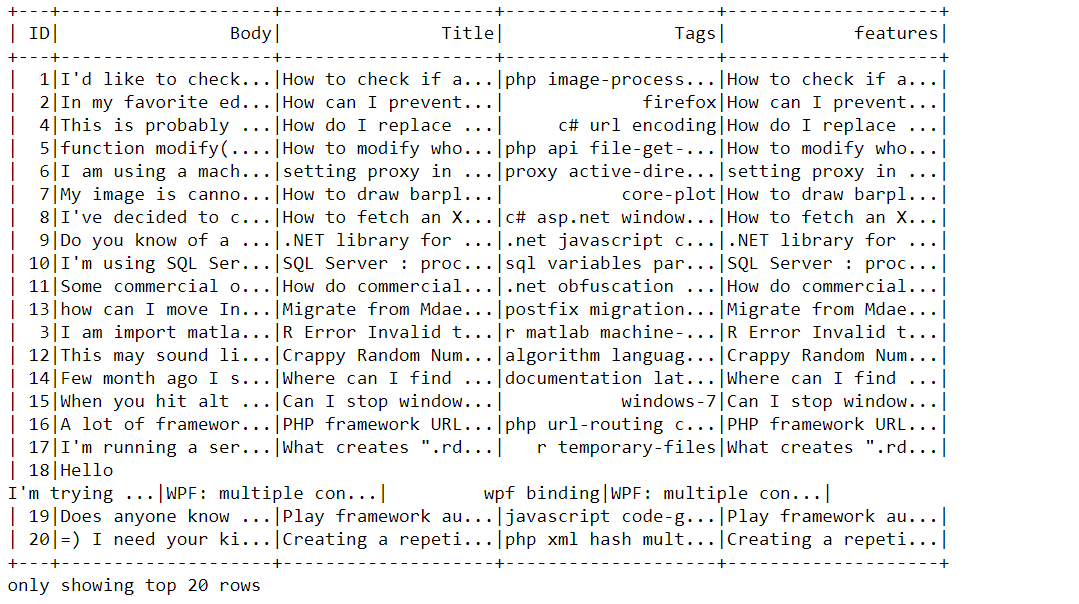
In the previous Facebook challenge, I have utilized “**OneVSRest**” classifier with LinearSVC algorithm to multi-label and multi-classify our data. In this, assignment, I will try to explore the clustering algorithm to predict the topic distribution for the StackOverflow, StackExchange and MathOverflow posts.

In this Project, I have explored the “Latent Dirichlet Allocation” algorithm , which takes into consideration the multivariate normal distributions of the data points.

**Step:1 Data Cleaning**

In the first step, I have merged the Body and the Title column to create a single column called as Features.



Then, I have performed data cleaning by removing the HTML tags and stop words using the ***BeautifulSoup*** library and the ***StopWordRemover*** transformed in PySpark.

**Step 2: Feature Transformation**

In the second, step, I have utilized the transformers *StringTokenizers* and *CountVectorizer* to subsequently tokenize and vectorize the feature column. Further, I have extracted the vocabulary of the text by using the countVectorizer.

Step 3: **Topic Modelling using Latent Dirichlet Algorithm**

In, this step I have utilized the LDA algorithm from the Spark library to segregate the data into group of 10 topic. In , the process, I have assigned a limit of 20 words/tags to each topic.

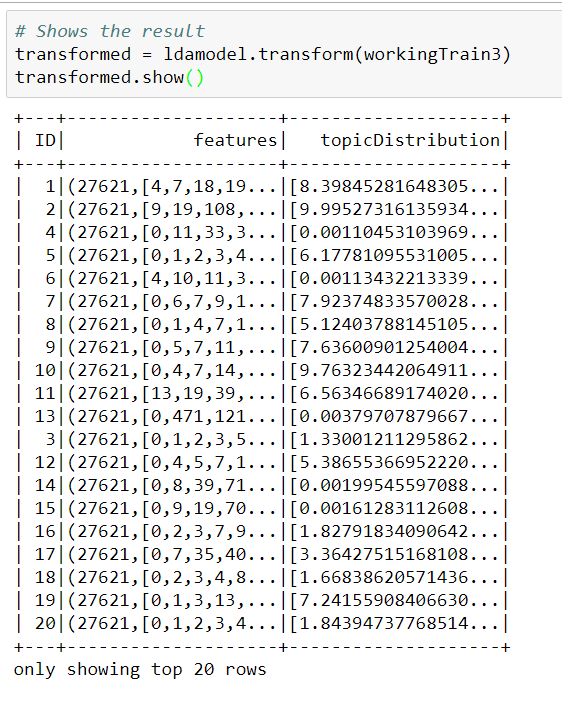
Words/Tags which are like each other are expected to cluster in the same group.

The below screenshot gives a glimpse of the same

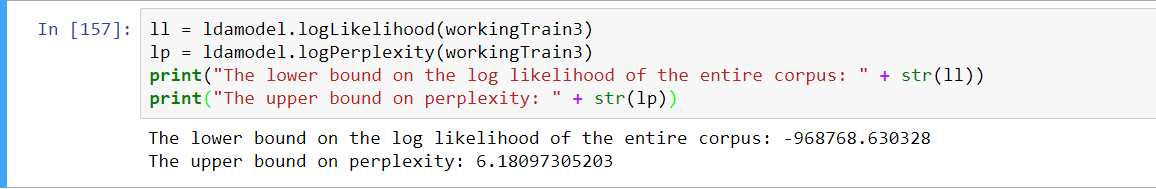


If we further increase the size of parameter K, we would be more efficiently able to predict the topic.

At, last, I have transformed my model using the LDA model, to show the probability of topic distribution for each stack exchange post.



**Step 5: Model Metrics**



Keeping in mind the infrastructure and other resource concern , I am limiting my analysis only on 1000 records of data.

Further:

I have additionally also experimented clustering using K-means and Gaussian Mixture Model, However GMM is facing outOfMemory issues.